Дата последней сборки документа:
28 ноября 2022 г. в 02:16

Конспект по книге Гудфеллоу «Глубокое обучение»*

Содержание

1. Численные методы

2. Основы машинного обучения

2.1. Точечная оценка

Точечное оценивание – это попытка найти единственное «наилучшее» предсказание интересующей величины. Пусть $\{x^{(1)}, \ldots, x^{(m)}\}$ – множество m независимых и одинаково распределенных точек. *Точечной оценкой*, или *статистикой*, называется любая функция этих данных

$$\theta_m = g(x^{(1)}, \ldots, x^{(m)}).$$

В этом определении не требуется, чтобы g возвращала значение, близкое к истинному значению θ, ни даже чтобы область значений g совпадала со множеством допустимых значений θ.

Алгоритм k-групповой перекрестной проверки применяется для оценивания ошибки обобщения алгоритма обучения A, когда имеющийся набор данных $\mathbb{D}$ *слишком мал* для того, чтобы простое разделение на обучающий и тестовый или обучающий и контрольный наборы могло дать точную оченку ошибки обобщения, поскольку среднее значение потери L на малом тестовом наборе может иметь высокую дисперсию.

2.2. Смещение

Смещение оценки определяется следующим образом

$$\mathrm{bias}(\hat{\theta}_m) = \mathbb{E}(\hat{\theta}_m) - \theta,$$

*Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение. – М.: ДМК Пресс, 2018. – 652 с.

где математической ожидание вычисляется по данным (рассматривается как выборка из случайной величины), а θ – истинное значение параметра, которое определяет порождающее распределение.

Оценка $\hat{\theta}$ называется *несмещенной*, если

$$(\hat{\theta}_m) = 0, \text{ т.е. } \mathbb{E}(\hat{\theta}_m) = \theta.$$

Оценка $\hat{\theta}_m$ называется *асимптотически несмещенной*, если

$$\lim_{m\to\infty} \operatorname{bias}(\hat{\theta}_m) = 0, \text{ т.е. } \lim_{m\to\infty} \mathbb{E}(\hat{\theta}_m) = \theta.$$

2.3. Дисперсия

Для определения смещения мы вычисляли математичесвкое ожидание оценки, но точно так же можем вычислить и ее дисперсию. *Дисперсией оценки* называется выражение

$$\operatorname{Var}(\hat{\theta}).$$

Стандартной ошибкой $\operatorname{SE}(\hat{\theta})$ называется квадратный корень из дисперсии.

Воспользовавшись центральной предельной теоремой, согласно которой среднее имеет приблизительно нормальное распределение, можем применить стандартную ошибку для вычисления вероятности того, что истинное математическое ожидание находится в выбранном интервале. Например, *95-процентный доверительный интервал* вокруг выборочного среднего (вокру оценки) $\hat{\mu}_m = \dfrac{1}{m}\sum\limits_{k=1}^{n} x^{(i)}$ определяется формулой

$$(\hat{\mu}_m - 1.96\,\operatorname{SE}(\hat{\mu}_m), \hat{\mu}_m + 1.96\,\operatorname{SE}(\hat{\mu}_m))$$

при нормальном распределении со средним $\hat{\mu}_m$ и дисперсией $\operatorname{SE}(\hat{\mu}_m)^2$.

NB: В экспериментах по машинному обучению принято говорить, что алгоритм A лучше алгоритма B, если верхняя граница 95-процентного доверительного интервала для ошибки алгоритма A меньше нижней границы 95-процентного доверительного интервала для ошибки алгоритма B.

2.4. Поиск компромисса между смещением и дисперсией для минимизации среднеквадратической ошибки

Что, если имеются две оценки, у одной из которых больше смещение, а у другой дисперсия? Какую выбрать?

Самый распространенный подход к выбору компромиссного решения – воспользоваться *перекрестной проверкой*. Эмпирически продемонстрировано, что перекрестная проверка дает отличные результаты во многих реальных задачах.

Можно также сравнить среднекваратическую ошибку (MSE) обеих оценок

$$\mathrm{MSE} = \mathbb{E}[(\hat{\theta}_m - \theta)^2] = \operatorname{bias}(\hat{\theta}_m)^2 + \operatorname{Var}(\hat{\theta}_m)$$

Желательной является оценка с малой MSE, именно такие оценки держат под контролем и смещение, и дисперсию. Соотношение между смещением и дисперсией тесно связано с возникающими в машинном обучении понятиями емкости модели, недообучения и переобучения.

Если ошибка обобщения измеряется посредством MSE (и тогда смещение и дисперсия становятся важными компонентами ошибки обобщения), то увеличение емкости (то есть *усложнение модели*) влечет за собой *повышение дисперсии* и *снижение смещения*.

2.5. Состоятельность

Обычно нас интересует также поведение оценки по мере роста размера обучающего набора. В частности, мы хотим, чтобы при увеличении числа примеров точечные оценки сходились к истинным значениям соответствующих параметров.

Формально это записывается в виде (условие состоятельности)

$$\hat{\theta}_m \xrightarrow{\mathbf{P}} \theta,\ (m \to \infty)$$

Иногда это условие называют *слабой состоятельностью*, понимая под *сильной состоятельностью* сходимость *почти наверное* $\hat{\theta}$ к θ.

Состоятельность гарантирует, что смещение оценки уменьшается с ростом числа примеров. Однако обратное неверно – из асимптотической несмещенности не вытекает состоятельность. Рассмотрим, к примеру, оценивание среднего μ нормального распределения $N(x;\mu,\sigma^2)$ по набору данных, содержащему m примеров: $\{x^{(1)},\dots,x^{(m)}\}$.

Можно было бы взять в качестве оценки первый пример: $\hat{\theta} = x^{(i)}$. В таком случае $\mathbb{E}(\hat{\theta})_m = \theta$, поэтому оценка является несмещенной вне зависимости от того, сколько примеров мы видели. Отсюда, конечно, следует, что оценка асимптотически несмещенная. Но она не является состоятельной, т.к. *неверно*, что $\hat{\theta}_m \to \theta,\ (m \to \infty)$.

2.6. Оценка максимального правдоподобия

Рассмотрим множества m примеров $\mathbb{X} = \{x^{(1)},\dots,x^{(m)}\}$, независимо выбираемых из неизвестного порождающего распределения $p_{data}(x)$.

Обозначим $p_{model}(x;\theta)$ параметрическое семейство распределений вероятности над одним и тем же пространством, индексированное параметром θ.

Тогда оценка максимального правдоподобия для θ определяется формулой

$$\theta_{ML} = \arg\max_{\theta} p_{model}(\mathbb{X};\theta) = \arg\max_{\theta} \prod_{i=1}^{m} p_{model}(x^{(i)};\theta)$$

Такое произведение большого числа вероятностей по ряду причин может быть неудобно. Например, оно подвержено *потере значимости*. Для получения эквивалентной, но более удобной задачи оптимизации заметим, что взятие логарифма правдоподобия не изменяет $\arg\max$, но преобразует произведение в сумму

$$\theta_{ML} = \arg\max_{\theta} \sum_{i=1}^{m} \log p_{model}(x^{(i)};\theta)$$

Поскольку $\arg\max$ не изменяется при умножении функции стоимости на константу, мы можем разделить правую часть на m и получить выражение в виде математического ожидания относительно эмпирического распределения $\hat{p}_{data}$, определяемого обучающими данными

$$\theta_{ML} = \arg\max_{\theta} \mathbb{E}_{x\sim\hat{p}_{data}}[\,\log p_{model}(x;\theta)\,]$$

Один из способов интерпретации оценки максимального правдоподобия состоит в том, чтобы рассматривать ее как минимизацию дивергенции (расхождения) Кульбака-Лейблера между этими эмпирическим распределением $\hat{p}_{data}$, определяемым обучающим набором, и модельным распределением.

Дивергенция Кульбака-Лейблера определяется формулой

$$D_{KL}(\hat{p}_{data}\,||\,p_{model}) = \mathbb{E}_{x\sim\hat{p}_{data}}[\,\log\hat{p}_{data}(x) - \log p_{model}(x)\,]$$

Первый член разности в квадратных скобках зависит только от порождающего данные процесса, но не от модели. Следовательно, при обучении модели, минимизирующей дивергенцию КЛ, мы должны минимизировать только величину

$$-\mathbb{E}_{x\sim\hat{p}_{data}}[\,\log p_{model}(x)\,],$$

а это, конечно, то же самое, что максимизация величины $\theta_{ML} = \arg\max_{\theta} \mathbb{E}_{x\sim\hat{p}_{data}}[\,\log p_{model}(x;\theta)\,]$.

NB: То есть, другими словами задача максимизации правдоподобия эквивалентна задаче минимизации дивергенции Кульбака-Лейблера между эмпирическим распределением $\hat{p}_{data}$ и модельным распределением p_{model}.

Список литературы

1. *Рамальо Л.* Python – к вершинам мастерства: Лаконичное и эффективное программирование. – М.: МК Пресс, 2022. – 898 с.
2. *Хейдт М., Груздев А.* Изучаем pandas. – М.: ДМК Пресс, 2019. – 682 с.